The Planetary Data System A Case Study in the Development and Management of Meta-Data for a Scientific Digital Library

J. Steven Hughes
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109
Steve.Hughes@jpl.nasa.gov

Abstract

The Planetary Data System (PDS) is an active science data archive managed by scientists for NASA's planetary science community. With the advent of the World Wide Web the majority of the archive has been placed on-line as a science digital library for access by scientists, the educational community, and the general public. The meta-data in this archive, originally collected to ensure that future scientists would be able to understand the context within which the science data was collected and archived, has enabled the development of sophisticated on-line interfaces. The success of this effort is primarily due to the development of a standards architecture based on a formal model of the planetary science domain. A peer review process for validating the meta-data and the science data has been critical in maintaining a consistent archive. In support of new digital library research initiatives, the PDS functions as a case study in the development and management of meta-data for science digital libraries. In addition the PDS looks forward to participating in digital library research areas, including interoperability and standard protocols.

1.0 Introduction

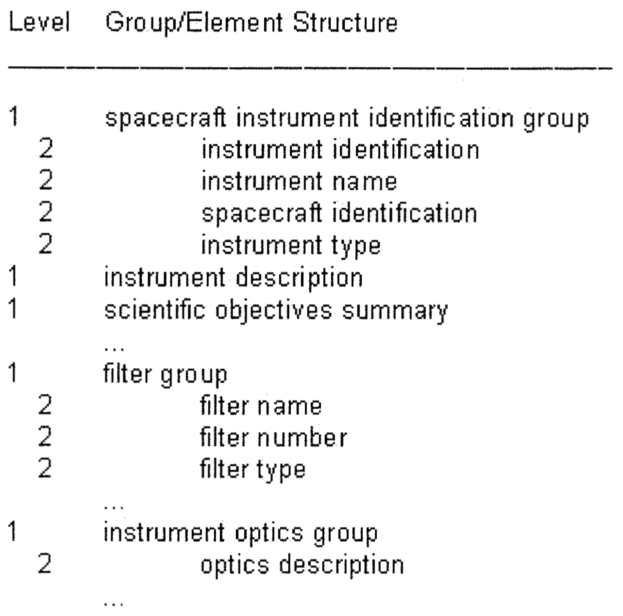
The Planetary Data System (PDS) [1] is an active science data archive managed by scientists for NASA's planetary science community that has been in operation since March, 1990. Envisioned as a long term archive, the PDS early on emphasized the development of a standards architecture that would include both the science data and the meta-data necessary for understanding the context under which the data were captured as well as interpreting diverse storage formats. This standards architecture includes a formal model of the planetary science domain, a standard grammar for encoding the information, and a standard language represented in the Planetary Science Data Dictionary. This standards architecture has been used to create a high quality science data archive of about five terabytes that is distributed on Compact Disk (CD) media. The meta-data in this archive, even though collected to ensure the usability of the science data for future scientists, has also allowed the majority of the archive to be made available through the World Wide WEB (WEB) as a digital library. The implementation of this archive as an on-line digital library provides an instructive case study in the development and management of meta-data for digital libraries.

In the following we will give a brief history of the PDS covering the early development of the standards architecture, the meta-data model, and describe how the meta-data in the archive has had a significant positive impact on the ability to support on-line search and access via the Web. In addition several "lessons learned" regarding the development and management of meta-data will be discussed.

2.0 Overview

In 1986, the Committee on Data Management and Computing (CODMAC) issued a report [2] that explored management approaches and technology developments for computation and data management systems designed to meet future needs in the space sciences. This report had resulted from earlier observations that a wealth of science data would ultimately cease to be useful and probably lost if a process was not developed to ensure that the science data were properly archived. In particular it was proposed that the data be transferred to stable media and that sufficient meta-data be included to ensure that future users of the data would be able to interpret the data as well as understand the context under which the data were collected.

After the development of a successful prototype, the PDS was funded in 1987 and work started on modeling the entities within the planetary science domain. Using formal modeling techniques, the team developed data structure charts that described in detail the data sets within the planetary science community and other related entities including missions, spacecraft, instruments, targets bodies, measured parameters, and bibliographic references. For example, in Figure 1, the instrument entity had sufficient detail so that a user of the system would have a good understanding of the instrument's operation without having to go to an instrument design document. If more detail was needed, the model allowed references to supporting documents using bibliographic citations. An example of a science data set is the collection of about 50,000 Mars images returned by the Viking Orbiter spacecraft in 1976. An individual image within this data set is called a data set granule.



Level	Group/Element Structure
1	spacecraft instrument identification group
2	instrument identification
2	instrument name
2	spacecraft identification
2	instrument type
1	instrument description
1	scientific objectives summary
	...
1	filter group
2	filter name
2	filter number
2	filter type
	...
1	instrument optics group
2	optics description
	...

Figure 1. Data Structure Chart for Science Instrument

An additional design goal was to allow sophisticated searches for the data through catalogs. In particular the scientists wanted the capability to find data sets through relationships with other entities. For example, using relationships between data sets, spacecraft, and instruments, scientists wanted the ability identify the images that had been captured using a specific filter, on a specific camera type, on a specific spacecraft. A simplified entity model showing these relationships is given in Figure 2.

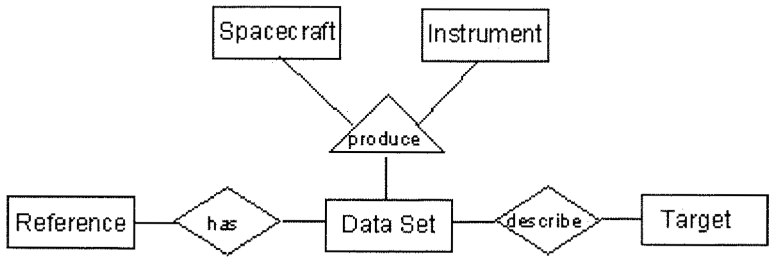
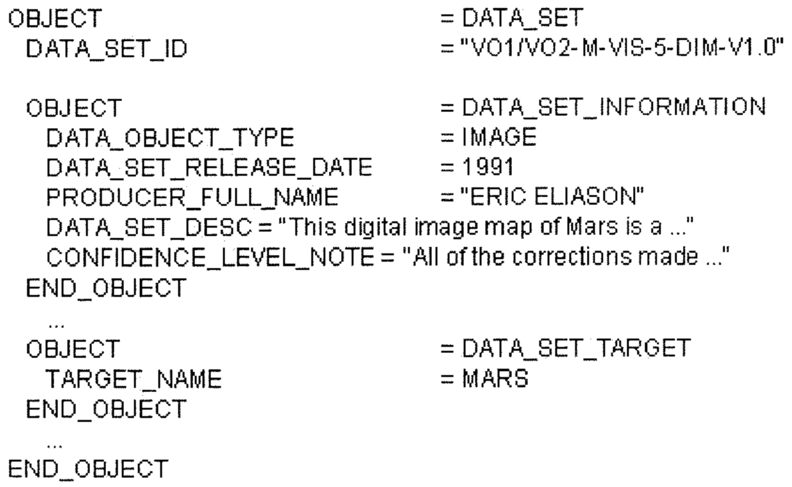


Figure 2. Simplified Entity Model

The PDS went on-line in 1990 with about 75 data sets in its archive. A high level data set catalog allowed the searching and ordering of any data set in the archive through an on-line interface with 88 user views. To bring data into the archive, the PDS developed a data ingestion procedure that included a formal peer review process. The peer review committee included peer scientists who reviewed the science data and collected meta-data for validity and usability while technical staff members reviewed the data for adherence to the standards architecture. The PDS currently has about 350 peer reviewed data sets in its archive with another 150 waiting the completion of peer review. The standards architecture and procedures, including the peer review process, are documented in three volumes, the Data Preparation Workbook [3], the Standards Reference [4], and the Planetary Science Data Dictionary (PSDD) [5]. The heart of the standards architecture was and continues to be the Planetary Science Data Dictionary.

3.0 The Data Dictionary

The original data structure charts used to model the entities in the planetary science domain were the basis for what was to become the Planetary Science Data Dictionary (PSDD). In a committee of scientists and technical staff, each attribute (element) that described an entity was used to define a data element in the data dictionary. For each attribute, the team created a name that conformed to a nomenclature standard, wrote a definition, assigned a data type, and either identified a range of values for numerical attributes or collected a set of possible values for enumerated attributes. They also created object classes by grouping the data elements by entity. Figure 3 shows the meta-data collected for a Mars image data set using the Data Set Object.



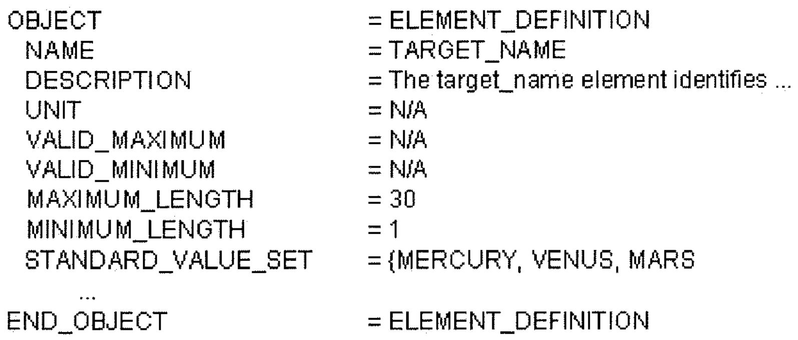
```
OBJECT                              = DATA_SET
  DATA_SET_ID                       = "VO1/VO2-M-VIS-5-DIM-V1.0"

  OBJECT                            = DATA_SET_INFORMATION
    DATA_OBJECT_TYPE                = IMAGE
    DATA_SET_RELEASE_DATE           = 1991
    PRODUCER_FULL_NAME              = "ERIC ELIASON"
    DATA_SET_DESC = "This digital image map of Mars is a ..."
    CONFIDENCE_LEVEL_NOTE = "All of the corrections made ..."
  END_OBJECT
    ...
  OBJECT                            = DATA_SET_TARGET
    TARGET_NAME                     = MARS
  END_OBJECT
    ...
END_OBJECT
```

Figure 3. Data Set Object

The initial release of the PSDD contained 455 data elements. It should be noted that where the development of the original data structure charts focused on detailed specification of attributes for each entity, the development of the data dictionary focused on generalization where similar attributes from different entities were merged into one data element. The level of generalization was continually debated in the committee with scientists typically arguing toward specification and data modelers arguing toward generalization.

The PSDD continues to grow as new data sets are ingested into the archive. In particular, a new instrument typically requires a new set of data elements to describe the resulting data. The process of adding data elements for new instruments and the resulting data is viewed positively since the scientists or his team is able to define the domain of discourse for all similar instruments and data that might be archived in the future. However there is a need to control the wholesale proliferation of new data elements. Typically a committee will first determine what existing data elements can be used, sometimes after proposing minor changes to the definition. The attempt is to define new data elements for only truly new characteristics or concepts. For example, the Viking Orbiter camera images were described with about 35 data elements. The Voyager camera images also required about 35 data elements, the majority of which had been used by Viking. The Galileo camera images however used the Voyager data elements and added over 60 more. As a more extreme example, the Magellan Synthetic Aperture Radar (SAR) used relatively few existing data elements and required the addition of over 100 new data elements. Figure 4 shows a meta-meta-data object used for ingesting new data elements into the data dictionary.



```
OBJECT                    = ELEMENT_DEFINITION
  NAME                    = TARGET_NAME
  DESCRIPTION             = The target_name element identifies ...
  UNIT                    = N/A
  VALID_MAXIMUM           = N/A
  VALID_MINIMUM           = N/A
  MAXIMUM_LENGTH          = 30
  MINIMUM_LENGTH          = 1
  STANDARD_VALUE_SET      = {MERCURY, VENUS, MARS
  ...
END_OBJECT                = ELEMENT_DEFINITION
```

Figure 4. Element_Definition Object

The PSDD currently contains over 1075 data elements and provides the primary domain of discourse for the Planetary Science Community. For example, before the PDS, the term spacecraft_id could have been a number, acronym - VO1, or a name - Viking Orbiter 1. It is now clearly defined as an acronym. In addition, the definition of the data element target_name with a standard values set has limited the number of new aliases for target bodies. The PDS also adheres to standards set by international standards organizations. The International Astronomical Union (IAU) gazetteer is used as the standard for planetary feature names and the Systeme Internationale table of standard units is used as a basis for units of measurement.

4.0 The Data Model

As described previously, a formal model of the planetary science domain resulted from the original data structure charts that described the entities within the planetary science community [6]. This model, described in the PSDD, was easily translated to the relational model for subsequent implementation in a relational database management system. The resulting database supported very complex searches for data sets. For example, a scientist could query for data sets that had been created using selected filters and detectors on specific instruments at a specific time.

It soon became apparent however that the task of collecting the data necessary to populate this model was problematic for two primary reasons. The first was that the scientists soon felt that the effort needed for collecting and organizing the information as required by the model was not worth the benefit of being able to find the data using sophisticated queries. This reasoning was supported by the then valid argument that the majority of the science users would not use the catalog to find the data in the first place. Most were familiar with the community and they knew where to go or who to ask to find the data. Any users unfamiliar with the community could be referred to a knowledgeable individual for help.

The second reason was that in certain areas the model was too rigid and complex. In particular, since a "black box" instrument model had been developed based on the relatively few instruments known at the time, new types of instruments did not always fit the model well. As a simple example, cameras have an optics section where magnetometers have none. In using the instrument model to describe a magnetometer, the optics section of the model had to be NULLed out. Situations such as this led many to feel that the instrument model was too camera specific.

Realizing that a meta-data collection and model-fitting bottle-neck existed, that the existing users did not need a sophisticated search capability, and that the volume of data to be ingested into the archive was rapidly increasing, the PDS decided to streamlined the model for the data set catalog. In particular, much of the information that had been captured as discrete data elements was aggregated into simple text descriptions. For example, the ten data elements that described the optics section of an instrument were eliminated. If the instrument had optics, they were described in a paragraph in the instrument description. However, data elements such as instrument_name, instrument_type, and instrument_desc that identify or describe the instrument were retained. In addition data elements that show a relationship to other entities were also retained. For example, instrument_host_id is used in the instrument model to link the instrument to the instrument host (spacecraft).

It is important to point out that the streamlining only occurred at the high- or data set level in the model. Specifically the mission, spacecraft, instrument, and data set models were streamlined to use more text, greatly simplifying the meta-data collection and manipulation for these entities. In regard to the requirement that sufficient meta-data exist for future users to understand the science context, the peer review process was modified to ensure that the eliminated data element information and additional references to supporting documents were included in the textual descriptions. This meta-data continued to be included on the archive media.

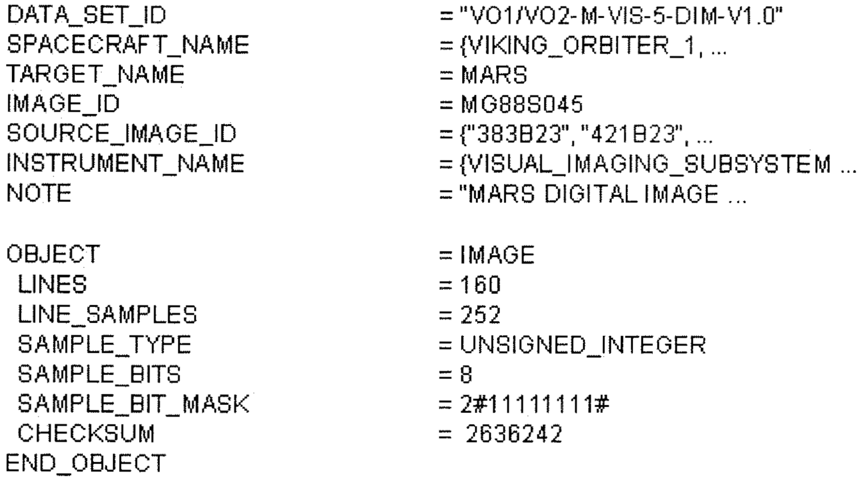
At the detailed- or individual granule level however, the model remained relatively complex. In fact, the amount of meta-data captured at the detailed level has significantly increased as data from more complex instrumentation has been ingested. To enable this increase in complexity, the meta-data model at this level has been kept flexible, allowing the addition of data elements as needed. The Galileo image example previously mentioned is a good example of the need for additional keywords to handle basically similar but more complex instrumentation and processing. This flexibility however necessarily complicates the development of generic catalogs for searching data sets at the granule level.

5.0 The Language

During the design phase of the PDS, the model for the planetary science community had been captured in a data dictionary. The need for a language to represent the model and to capture meta-data for the archive resulted in the development of the Object Description Language (ODL), a language consisting of "keyword = value" (keyword/value) statements. The primary requirements for this language were flexibility, simplicity in meta-data representation, and readable by humans as well as machines. Each keyword represents an attribute (element) in the original data structure charts and was defined as a data element in the data dictionary.

The need for the aggregation of keyword/value statements to describe entities resulted in the addition of a grouping mechanism where the statements were grouped into objects bracketed by "OBJECT=entity" and "END_OBJECT" statements. Using this capability, an object class was created for each entity in the model. These were included as a core part of the PDS standards architecture. A formal grammar was defined for the language and several language parsers were developed.

A mandatory requirement for the ingestion of a data set into the PDS archive is that the meta-data describing the data set, its component granules and the associated spacecraft, instrument, mission, and other related entities be captured in ODL using the standard models. This information is then written into ASCII text files called labels and written with the data onto the archive volume which is typically Compact Disk (CD) media. Once an archive volume has been reviewed and archived, the meta-data can be extracted for use in catalogs, inventories, other specialized search aids. Figures 3 and 5 illustrate portions of labels for an image data set and one of its images respectively.



```
DATA_SET_ID                         = "VO1/VO2-M-VIS-5-DIM-V1.0"
SPACECRAFT_NAME                     = {VIKING_ORBITER_1, ...
TARGET_NAME                         = MARS
IMAGE_ID                            = MG88S045
SOURCE_IMAGE_ID                     = {"383B23","421B23", ...
INSTRUMENT_NAME                     = {VISUAL_IMAGING_SUBSYSTEM ...
NOTE                                = "MARS DIGITAL IMAGE ...

OBJECT                              = IMAGE
 LINES                              = 160
 LINE_SAMPLES                       = 252
 SAMPLE_TYPE                        = UNSIGNED_INTEGER
 SAMPLE_BITS                        = 8
 SAMPLE_BIT_MASK                    = 2#11111111#
 CHECKSUM                           =  2636242
END_OBJECT
```

Figure 5. Image Label

The PDS has found that translation of ODL to other languages has been relatively easy. In particular, meta-data in ODL is translated to SQL insert statements to load a relational database for the data set catalog. In addition, being object-based, meta-data in ODL has been readily converted to more formal object-oriented representations.

6.0 Categories of Meta-data

Early on the PDS determined the need to capture many categories of meta-data for the requirements of a long term archive. Although not clearly identified in the early design of the system, the continued use of the meta-data in the archive has necessarily resulted in an informal categorization of the meta-data collected. These have been loosely described as structural and catalog.

6.1 Structural Meta-Data

Structural meta-data typically refers to information about how the entity is represented on the archive media. Within the structural category, there are roughly three subcategories. The first of these is that required to understand the data representation as written by the hardware on the media. For example, the PDS is often required to archive data for machine architectures that no longer exist and for which there are no resources for conversion. In such cases the raw data are simply transferred to the archive media and described sufficiently so that future users can successfully interpret the raw data. An example is the use of Binary Coded Decimal (BCD), a format that is seldom used for scientific data but which was found in an older data set. Second and closely related is meta-data typically associated with a computer's file system such as file name, record type, and record size. Finally, there is the meta-data needed to describe the structure of the data as organized by the data producer. For example, a typical Viking Orbiter image is stored as a raster image of 800 lines by 800 line_samples of 8 bits each.

6.2 Catalog Meta-Data

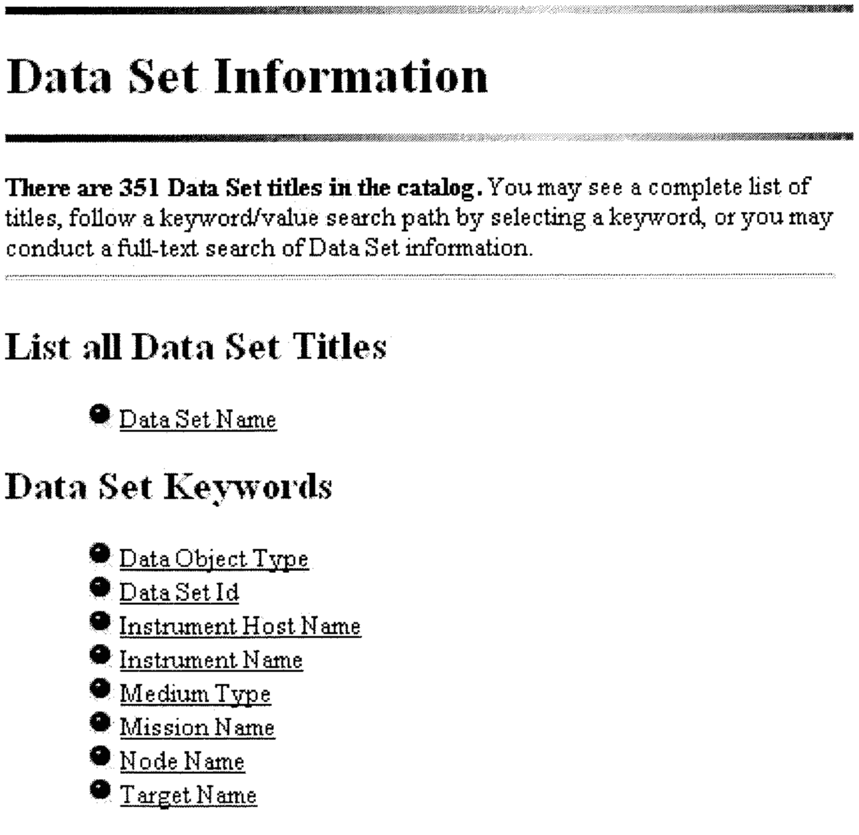
The catalog category includes meta-data useful for identifying the data or describing the context within which the data were captured. It is especially useful for building catalogs and inventories. Within this category there are again three subcategories. First, "identification" meta-data is used to uniquely identify the object being described as well as to identity other entities that are related. For example, a Viking Orbiter image will have a unique image_id, the data_set_id of the data set that the image is contained in, and the instrument_id of the instrument that captured the image. Second, "description" meta-data provides information useful for understanding the entity. For example, the Viking Orbiter image includes exposure_duration and filter_name as data elements to describe the context within which the image was taken. Notice that all these keywords are actually detailed instrument attributes which are included as part of the image description since their values are specific to the image. Finally, other "property" meta-data may be collected to describe modeled aspects of a data set granule. For instance, if an image is a part of a digital map, then a map_projection object is used to group map attributes such as projection_type and resolution.

7.0 Lessons Learned

7.1 The Importance of a Model

The existence of a formal model for the planetary science domain was critical for the development of a long term science data archive. As mentioned previously, the primary reason for the model was to support the collection of meta-data for the archive so that future users would know the context within which the data were taken.

The availability of a formal model also made the development of sophisticated on-line interfaces relatively easy. Without such as model, search capabilities would have been essentially limited to free text searches. Figure 6 shows an example of an early Web interface to the data set catalog. This interface provides listings of data sets grouped by selected keywords. By clicking on the link, a dynamic list of pertinent data sets is generated.



Data Set Information

There are 351 Data Set titles in the catalog. You may see a complete list of titles, follow a keyword/value search path by selecting a keyword, or you may conduct a full-text search of Data Set information.

List all Data Set Titles

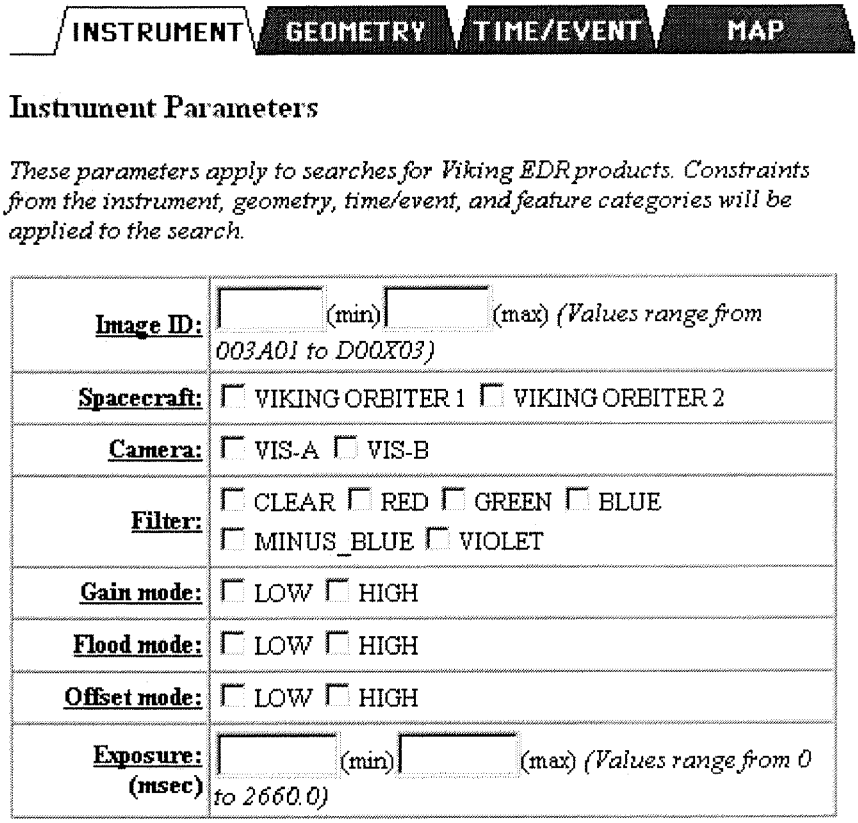
- Data Set Name

Data Set Keywords

- Data Object Type
- Data Set Id
- Instrument Host Name
- Instrument Name
- Medium Type
- Mission Name
- Node Name
- Target Name

Figure 6. Data Set Catalog Interface

At the detailed- or data set granule level, comprehensive models for the data were developed using all the information available and new attributes were added as needed. Again even though the primary purpose of the meta-data was to support future users, interface developers never complained about too much meta-data being available for searching purposes. Figure 7 shows an example of an interface to the Viking image catalog. The user enters constraints for the search and the system responds with a list of matching browse images, image attributes, and selection buttons for ordering.



Instrument Parameters

These parameters apply to searches for Viking EDR products. Constraints from the instrument, geometry, time/event, and feature categories will be applied to the search.

Image ID:	(min) (max) *(Values range from 003A01 to D00X03)*
Spacecraft:	□ VIKING ORBITER 1 □ VIKING ORBITER 2
Camera:	□ VIS-A □ VIS-B
Filter:	□ CLEAR □ RED □ GREEN □ BLUE □ MINUS_BLUE □ VIOLET
Gain mode:	□ LOW □ HIGH
Flood mode:	□ LOW □ HIGH
Offset mode:	□ LOW □ HIGH
Exposure: **(msec)**	(min) (max) *(Values range from 0 to 2660.0)*

Figure 7.Viking Image Catalog Interface

Most important however, without the model the consistency of the meta-data could not have been maintained. The standard models guided the collection of the meta-data and allowed the development of software to perform syntactic and semantic validation. In particular the data dictionary was used by the software to validate data elements and their values.

7.2 The Model Fitting Problem

The most difficult part of developing a digital archive is by far the collection, model fitting, and validation of the meta-data. In fact a major part of the PDS budget is used to support this effort. Where the actual science data requires a certain level of effort for collection and validation, familiarity with the data makes this task manageable. The collection of meta-data however is typically a research task requiring access to a variety of information sources. In the PDS community this could be instrument designers, spacecraft developers, mission planners, navigation experts, software developers, as well as the principle investigators.

Once the data has been collected, it must be modified to fit the existing models. A simple example would be the collection of date-time information and converting it to a standard format to ensure consistency and ease of use, such as yyyy-mm-ddThh:mm:ss.sss. A more complex example would be the ingestion of data from a new source, such as a rover on the surface of a planet. Where the navigation models had been developed for instruments either in flight or stationary on a surface, a new navigation model had to be developed for an instrument moving on a surface. In summary, making the collected meta-data fit an existing model versus changing the model is the difficult tradeoff constantly encountered in ingestion.

7.3 Meta-Data Validation

The PDS uses a formal peer review to validate both the meta-data as well as the science data. Focusing on meta-data, scientists check the meta-data for validity and usability and technical staff checks it for adherence to standards. After a peer review, an "in lien resolution" phase allows changes to the meta-data before it is accepted as part of the archive. The peer review process has always been considered a integral part of the ingestion process, ensuring the usefulness of the data. However it has also had a critical role in maintaining a consistent set of meta-data. It is readily apparent from the inconsistencies that have appeared in spite of peer review, that the meta-data in the archive would become essentially useless without it.

8.0 The PDS as a Digital Library

By its very nature as a digital archive containing both data and meta-data, the PDS is a digital library. Meta-data collected for the archive was loaded into a database and an on-line interface was developed to allow search and ordering of the data. Initially computer technology limited access to the digital library to a relatively small group of users.

As technology has progressed however, new and more powerful user interfaces have been developed. With the advent of the Web and the development of Web interfaces [7], the PDS found that its customer base grew drammatically. However the meta-data model changed little to meet the new interface requirements. The majority of changes were syntactical as opposed to semantic.

The PDS is a collection of federated, heterogeneous nodes that are distributed geographically. These nodes focus on different science disciplines and have implemented dissimilar data management systems locally. The PDS standards architecture alone forces a commonality across the system. Based on this architecture, a federation of search aids such as catalogs and inventories have been developed to give users the ability to search and access the entire archive without knowledge of its distributed nature. Figure 8 shows a hierarchy of the search aids available. At the highest level, the Distributed Inventory System (DIS) allows users to search for either archived or pending data sets or any resource that supports the use of the data. Twenty attributes are available for constraining the searches including data type, time, and related missions, instruments, spacecraft, and target bodies. At the data set level, the data set catalog allows the search and ordering of archived data sets. The detailed level has several sophisticated catalogs for individual granule searching, including map based search for images. If not available through a catalog interface, the majority of the remaining archive is available on-line either as CD volumes or in disk farms.

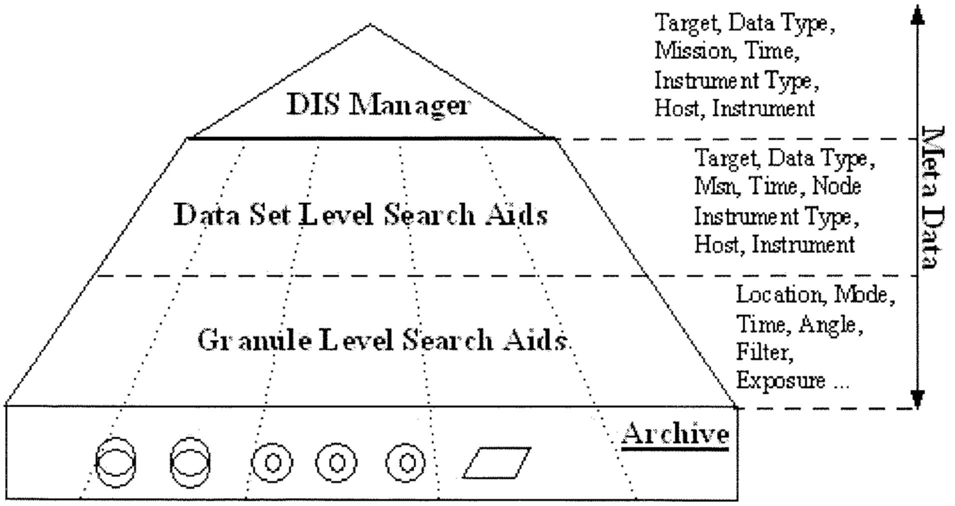


Figure 8. Search Aid Hierarchy

9.0 Current Research

The PDS is now interested in developing a digital library interface using standards and technology that has resulted from digital library research funding, in particular the Z39.50 standard protocol and the library (BIB-1) and Earth Sciences (GEO-1) profiles. We are interested in developing space science profile(s) to support interoperability across diverse and distributed science data archives as well as supporting access by the educational community and the general public. Cursory review of the problem has revealed that the meta-data currently in the PDS archive will readily allow a planetary science profile to be developed. It is interesting that at the data set level, the PDS has about ten data elements that have similar counterparts in other disciplines within the space sciences community. Several proposals are now being considered to research interoperability between these systems.

10.0 Conclusion

The PDS as a science data archive, developed a model of the planetary science domain that includes spacecraft, instruments, missions, target bodies, references, data sets and data set granules. It was developed using a formal methodology for modeling entities and their relationships within a domain. This model is a fundamental part of the PDS standards architecture and is described in the Planetary Science Data Dictionary. Its function is to capture the meta-data needed to understand the instruments that captured the science data, the formats used to store the data, and the mission objectives for getting the data, so that scientists will be able to intelligently use the data in the future. The meta-data is validated in a peer review and is extracted for catalogs, inventories, and other on-line search aids.

The existence of the formal model and the collected meta-data has allowed the PDS to easily develop on-line catalogs and access tools. The conversion of the meta-data from ODL format to other models and languages has been easy because of the existence of the model and the consistency of the meta-data. The hard problems continues to be the collection, organization, and validation of the meta-data. With the advent of the Web the PDS has placed the majority of the archive on-line resulting in a scientific digital library that supports the science community, the educational community, and the general public.

Given past experiences, the PDS is well situated for developing interfaces that allow access to the PDS archive via standardized protocols such as Z39.50, standard languages such as XML, and standard interfaces now being considered for globally accessible digital libraries.

Acknowledgements

The research described in this paper was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

The author would also like to acknowledge Susan K. McMahon, the PDS project manager, Michael D. Martin, the Data Distribution Laboratory director, and the PDS staff, both past and present who have all worked hard to made the PDS a model science data system.

References

1. Arvidson, R.E., Dueck, S.L., "The Planetary Data System", Remote Sensing Reviews, 1994, Vol.9, pp.255-269.
2. Arvidson, R.A., etal, Issues and Recommendations Associated with Distributed Computation and Data Management Systems for the Space Sciences, National Academy Press, 1986.
3. Planetary Data System Data Preparation Workbook, JPL Internal Document, JPL D-7669; Part 1, Jet Propulsion Laboratory, April 21, 1993. [Also accessible at http://pds.jpl.nasa.gov/prepare.html]
4. Planetary Data System Standards Reference, JPL Internal Document, JPL D-7669; Part 2, Jet Propulsion Laboratory, July 24, 1995. [Also accessible at http://pds.jpl.nasa.gov/prepare.html]
5. Planetary Science Data Dictionary Document, JPL Internal Document, JPL D-7116; Rev D, Jet Propulsion Laboratory. [Also accessible at http://pds.jpl.nasa.gov/prepare.html]
6. Hughes, J.S., Li, Y.P., "The Planetary Data System Data Model", Proceedings of Twelfth IEEE Symposium on Mass Storage Systems, April 25-29, 1993, pp 183-189.

7. Hughes, J.S., Bernath, A.M., "The Planetary Data System Web Catalog Interface - Another Use of the Planetary Data System Data Model", Proceedings of the Fourteenth IEEE Symposium on Mass Storage Systems, September 11-14, 1995, pp.263-273.